BREAST CANCER DIAGNOSIS USING MACHINE LEARNING

Table of Contents

1. INTRODUCTION 3
1.1 BACKGROUND 3
1.2 AIM 4
1.3 ETHICAL, SOCIAL, PROFESSIONAL, LEGAL AND SECURITY CONSIDERATION 5
2. DATA 5
2.1 Data Description 5
2.2 DATA PREPROCESING 9
2.3 Packages Used For Analysis 10
3. COMPARATIVE ANALYSIS 10
3.1 METHODOLOGY 10
3.2 MACHINE LEARNING ALGORITHMS 11
3.3 DATA PREPARATION 13
3.4 EVALUATION METRICS 13
3.5 HYPERPARAMETER SEARCH 14
4. RESULT 14
4.1 BASELINE MODEL RESULT 14
4.2 OPTIMISED MODEL RESULT 15
4.3 Model interpretability shap analysis 15
5 CONCLUSION 17
6 REFERENCES 17

1. INTRODUCTION

1.1 BACKGROUND

The most common cancer in women is breast cancer. While the majority of breast cancers are benign and treatable with surgery, 25% have a latent, sneaky nature that causes them to develop slowly but spread quickly. Present treatments greatly slow the growth of tumors, but recurrence is unavoidable and leads to high death rates. The genesis of breast cancer cell activity is where the seeds are sown. Mammary development is characterized by alterations in cell contact and movement of the invasive, motile embryonic mammary cells.

With almost 12,000 fatalities in 2019, breast cancer remained the leading cause of cancer-related mortality among women. Additionally, it is the leading cause of death for people of all ages: 28% before the age of 50, 21% between the ages of 50 and 69, and 14% beyond the age of 70 (GCO 2019). Without notable variations based on age or geography, the overall 5-year survival rate in Italy is around 87%; the overall 10-year survival rate is almost 80%. Two key variables have contributed to the recent improvements in breast cancer prognosis: advancements in therapy tailoring and early diagnosis, especially when the disease is still in the subclinical stage. (ISTAT)

Large-scale population screening programs using regular mammography have significantly decreased the death rate from breast cancer in western nations; recently, Italy has also shown a fall in breast cancer mortality along with an increase in incidence. Although the effectiveness of screening mammography is still up for question, a number of studies have shown the value of planned population programs, including the significance of an early diagnosis. Innovative technology tests enabled a better subtype categorization that allowed for novel targeted therapeutics in endocrine-sensitive and HER2-positive illnesses, coinciding with an increase in screening program compliance. In specialized clinical follow-up settings, one of the most difficult topics to address is breast cancer surviving. This is mostly because of the 20 years of increased survival, which has led to the chronicization of the illness in its advanced stages and its cure in its early stages. The most significant side effects of cancer therapies that call for careful monitoring and specialized resources have been covered in this overview. This is especially crucial because, in the modern world, breast-specific follow-up generally stopped 5–10 years following a diagnosis of breast cancer. However, more funding has to be set aside to manage "breast cancer survivorship" in order to improve the general quality of life for those who have survived breast cancer (Vachanaram at.al 2020).

A variety of data mining and machine learning methods (M. Bahaj, 2018) are being utilized to predict breast cancer. One of the key tasks is determining the most relevant and acceptable algorithm for breast cancer prediction.

The format of this document is as follows: The link between machine learning, data gathering, feature engineering, data preprocessing, and machine learning algorithms is covered in Section 2. In Section 3, the approach and dataset description are explained. A theoretical explanation of the outcome analysis, project assessment, and reflection can be found in Sections 4 and 5.

The process of extracting valuable information from large datasets is known as data mining. Data mining functions and techniques can be used to identify any type of disease. For example, machine

learning, statistics, databases, fuzzy sets, data warehouses, and neural networks can be used to diagnose and prognosticate a variety of cancer diseases, including leukemia (M Shaheen, 2012), prostate cancer, and lung cancer (D. Delen, 2009). The three tests that make up "the gold standard" technique of cancer diagnosis are radiological imaging, pathology testing, and clinical evaluation (A. Reddy 2020). While the latest machine learning approaches and algorithms are focused on model creation, the conventional method uses the regression process to detect the existence of cancer. According to Z. Salod (2019), the model is intended to forecast data that has not yet been observed and delivers the desired outcome throughout training and testing. Three primary tactics form the basis of the machine learning process: preprocessing, feature selection or extraction, and classification (H. Kutrani, 2019). The primary function of machine learning is feature extraction, which is helpful in the diagnosis and prognosis of cancer. It can differentiate between benign and malignant tumors (E. Frank 2005).

1.2 AIM

Breast cancer diagnosis in medicine also makes use of machine learning. Based on data from the Scopus databases, Fig.1 displays the statistics for machine learning and research on the classification and detection of breast cancer from 2010 to 2019.

In order to direct future directions that may result in increases in patient safety, the current study looked at past patterns in breast cancer treatment for diabetic patients admitted to various hospitals. In this case study, we examine how machine learning might help distinguish between tumors into malignant (cancerous) or benign(non cancerous)

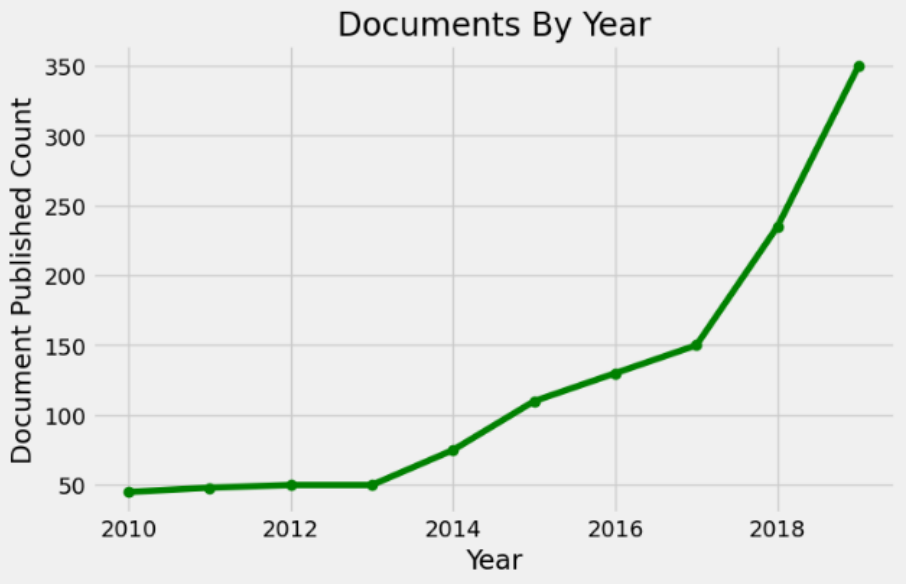


Figure 1: Histogram of machine learning for classification and detection of breast cancer publications.

1.3 ETHICAL, SOCIAL, PROFESSIONAL, LEGAL AND SECURITY CONSIDERATION

This study is based on a dataset regarding diabetes that was obtained via Kaggle. There is no personally identifiable information (PII) in the dataset. Since the records used in this study were anonymised, neither consent nor human subjects research was necessary.

2. DATA

2.1 Data Description

The Data Contained 32 Features where Thirty One (31) are the Input Features and One (1) Target column having 569 distinct rows which in total contains 31 Numerical features and 1 categorical features.

S/N	Feature Name	Type	Description	Nunique	Count
1	id	Integer	Unique identifier for each patient or data point in the dataset.	569	569
2	diagnosis	Categorical	contains the target variable, indicating whether the tumor is malignant (M) or benign (B).	2	569
3	radius_mean	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the mean radius of tumor.	456	569
4	texture_mean	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the mean texture of tumor.	479	569
5	perimeter_mean	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the mean perimeter of tumor.	522	569
6	area_mean	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the mean area of tumor.	539	569
7	smoothness_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of smoothness mean	474	569

8	compactness_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of compactness mean	537	569
9	concavity_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concavity mean	537	569
10	concave points_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concave points mean	542	569
11	symmetry_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of symmetry mean	432	569
12	fractal_dimension_mean	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of fractal dimension mean	499	569
13	radius_se	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Standard Deviation radius of tumor.	540	569
14	texture_se	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Standard Deviation texture of tumor.	519	569
15	perimeter_se	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Standard Deviation perimeter of tumor.	533	569
16	area_se	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Standard Deviation area of tumor.	528	569

17	smoothness_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of smoothness Standard Deviation	547	569
18	compactness_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of compactness mean	541	569
19	concavity_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concavity Standard Deviation	533	569
20	concave points_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concave points Standard Deviation	507	569
21	symmetry_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of symmetry Standard Deviation	498	569
22	fractal_dimension_se	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of fractal dimension Standard Deviation	545	569
23	radius_worst	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Worst (the mean of three largest values) radius of tumor.	457	569
24	texture_worst	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Worst (the mean of three largest values) texture of tumor.	511	569
25	perimeter_worst	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the	514	569

			Worst (the mean of three largest values) perimeter of tumor.		
26	area_worst	Integer	measurements or features derived from medical imaging (like mammograms or ultrasound) related to the Worst (the mean of three largest values) area of tumor.	544	569
27	smoothness_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of smoothness Worst (the mean of three largest values)	411	569
28	compactness_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of compactness Worst (the mean of three largest values)	529	569
29	concavity_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concavity Worst (the mean of three largest values)	539	569
30	concave points_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of concave points Worst (the mean of three largest values)	492	569
31	symmetry_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of symmetry Worst (the mean of three largest values)	500	569
32	fractal_dimension_worst	Integer	characteristics of cell nuclei present in the tumor, which can be indicative of cancerous growth of fractal dimension Worst (the mean of three largest values)	535	569

2.2 DATA PREPROCESING

Although minimal pretreatment was necessary, the data did not include noisy information as was anticipated. Data preprocessing is essential to clean the data, removing noise, inconsistent, and duplicated characteristics. Data duplication was checked for as it would lead to redundancy and worsen performance rather than enhance the computational complexity of the model. Certain characteristics were eliminated since they had unique identifiers (id, Unamed:32). As anticipated, additional preprocessing revealed that no feature had any missing data.

The majority of machine learning algorithms are built under the presumption that the training data is balanced, and they implicitly assume that the cost of each misclassification error is the same. However, this is not the case when working with real-world data, which is typically unbalanced (Huang, B., et al 2008). The problem of imbalance target data is pervasive and problematic; it hinders a large segment of the data community from developing predictive models.

Figure 2: Pie Chart Showing the Data Distribution of the Target

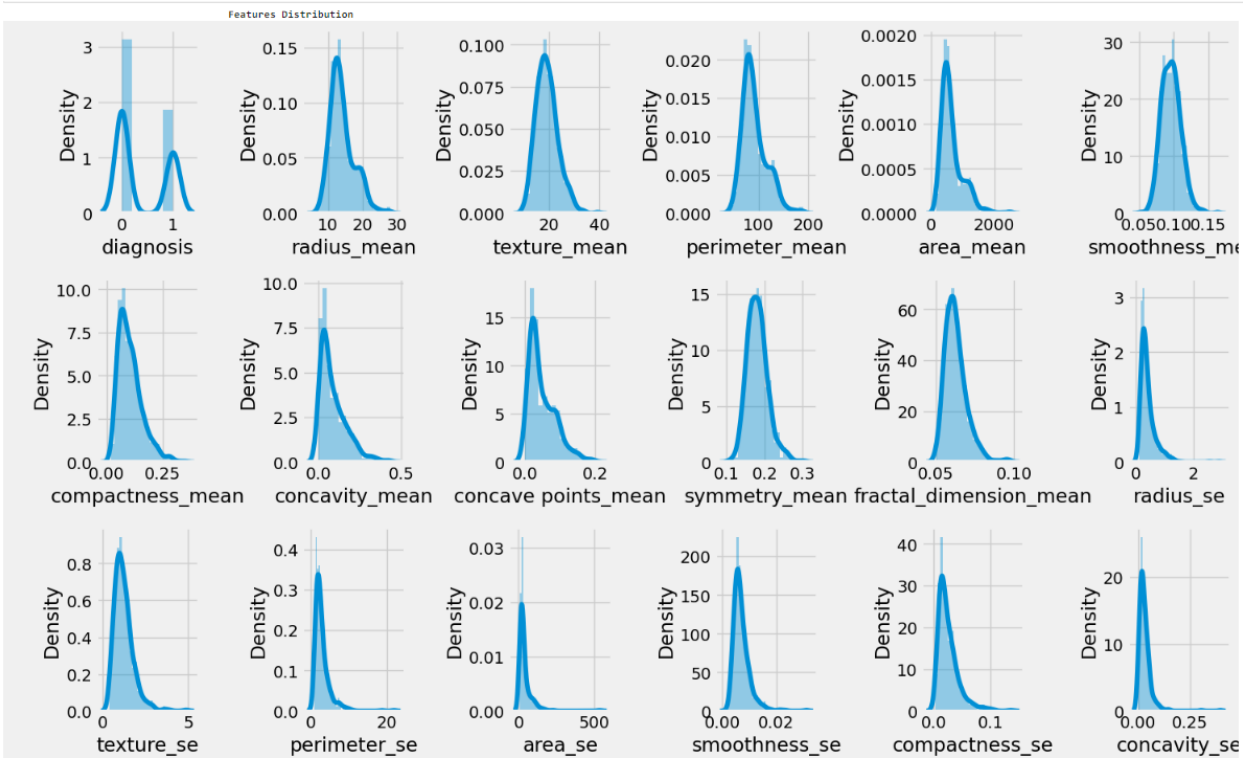


Figure 3: Distribution Plot for Different columns

2.3 Packages Used For Analysis

- **Pandas**: is an open-source Python Data Analysis Library, free to use (under a BSD licens e); it is one of the most popular and widely used data munging/wrangling tools. It gathers data (from a CSV or TSV file or a SQL database) and turns it into a Python data frame wi th rows and columns like a table in statistical applications such as Excel or SPSS (Bronsht ein Adi).

- **NumPy**: It is a Python library that includes a multi-dimensional array adding support for multi-dimensional arrays and matrices as well as several derivative objects (such as mask ed arrays and matrices, contains a variety of routines for performing quick array operation s, such as mathematical, logical, basic statistical operations, shape manipulation, selecting , sorting, I/O, random simulation, discrete Fourier transforms, and more (GeeksforGeeks, 2021)

- **Scikit-Learn**: Scikit-learn is a free machine learning library that focuses on data modellin g data by providing a consistent Python interface for supervised and unsupervised learnin g algorithms. It features many classifications, Regression, and clustering algorithms and i s designed to work with other scientific and numerical libraries, including NumPy (J. Bro wnlee 2021)

- **Scipy:** SciPy is a library of numerical routines for the Python programming language that provides fundamental building blocks for modeling and solving scientific problems. SciP y includes algorithms for optimization, integration, interpolation, eigenvalue problems, al gebraic equations, differential equations and many other classes of problems; it also provi des specialized data structures, such as sparse matrices and k-dimensional trees (Virtanen, P., 2020)

3. COMPARATIVE ANALYSIS

3.1 METHODOLOGY

As mentioned in 2.0, our goal in this research is to use an improved yet robust classification scheme from various machine learning classifiers that have been implemented/developed by various research teams, ranging from Lightgbm(Machado, M.R, 2019), Catboost (Dorogush, A.V, 2018), RandomForest (Breiman, L., 2001, Hackeling, G., 2017), Logistic Regression(Karp, A.H, 1998),Gradient Boosting Classifier(Bentéjac, C, 2021) to determine whether a patient has breast cancer or not based on the analysis and preprocessing done on the data. We conclude from the literature study that existing machine learning classifiers require extensive expert consultation for pre-processing and performance adjustment, and they perform quite well on related models. This procedure made use of Optuna (Akiba, T. 2019) since it has been shown to get the best parameter search result based on data and algorithm search parameter.

Variations in domains and hence in data necessitated that. Our use case is understanding patterns of breast cancer, and we want to develop a powerful classifier that is small and portable. To do

this, we first carry out an experimental research using default search parameters, serving as the baseline model. Next, we design a collaborative environment to generate an ensemble learning strategy. The main motivation for this research's experimental investigation—which seeks to determine which algorithm performs best for upcoming parameter training—is a comparative analysis of existing algorithms. We start our experiment by training a subset of classifiers across several rounds on a given dataset, which gives us an overview.

The methodology Implemented include

▪ The training phase is referred to as the assessment phase. Each potential classifier will go through a customized training and performance optimization process.

▪ Since we choose our algorithms using a range of factors, such as regression, tree-based models, and neighborhood models, our procedure does not specify the quantity of classification techniques. For example, the eight classification strategies we apply in this study might vary depending on the dataset's size and associated feature groups.

▪ Instead of using a percentage split, we use a stratified k-fold split because to the uneven class size of the target data (Khushi, M., 2021).

▪ Precision and Recall are used to create the F1 Score weighted average, which will be used to evaluate each possible classifier. The same dataset was utilized for both the ensemble and evaluation phases.

3.2 MACHINE LEARNING ALGORITHMS

Logistic Regression: According to (Armengol E. and De Mantaras, R.L.1998), When the response variable is discrete, logistic regression—a extension of linear regression—is utilized to estimate binary or multi-class dependent variables. Using sigmoid functions, a statistical model known as logistic regression models binary dependent variables. Sigmoid (logistic) functions are suitable for statistical investigations of binary classification issues because they consistently place any real number in the interval between 0 and 1. Logistic regression works similarly to linear regression in that it uses predefined weights or coefficients to calculate the output for a given input. The only difference is that logistic regression always yields binary outputs of 0 or 1. When classifying low-dimensional data with nonlinear bounds, logistic regression is utilized. For the problem of binary classification, as described in (Cao, C., 2012),

Decision Tree: Decision trees are modeled as a tree structure model, where leaves represent class labels and branches represent feature conjunctions. This model may be applied to classification tasks. One of the many supervised learning methods used to solve classification and regression problems is the decision tree technique. The main goal is to create a training model that can use the learnt decision rules to forecast the target variables. There are two types of decision tree

algorithms: continuous variable decision trees and categorical variable decision trees (Kögel, B. a. 2013), depending on the kind of target variables. A classification tree is the result of the learning process; each leaf in the tree represents a single value of the target variable, and the split at each node in the tree represents an if-then decision rule.

Gradient Boosting: A prediction model is built from an ensemble of weak prediction models using gradient boosting, a machine learning technique for classification and regression problems. This is because decision trees are commonly used in The poor learning process of decision trees, which usually beats random forest, results in gradient enhanced trees. Like other boosting models, it also uses wise step approaches to optimize a random differentiable loss function.

Random Forest: Random Forest is an ensemble learning technique for various tasks such as classification and regression. It starts with a large number of decision trees built during training and outputs a class that is the mean/average prediction (in the case of regression) or the mode of the classes (in the case of classification) of the individual trees (Tin Kam Ho 1995, Breiman, L., 2001). For regression tasks, the mean or average prediction provided by each individual tree is returned. Random decision forests counteract decision trees' propensity to overfit to their training set. Random forests outperform choice trees in most situations, while being less precise than gradient-boosted trees. However, the uniqueness of the data could affect how well they function.

Light Gradient Boosting Machine: LightGBM was another method developed by Microsoft Research Asia (Xie Y, 2018) that utilized the GBDT architecture. It makes an effort to boost computer effectiveness in order to more effectively handle big data prediction problems. An algorithm for tree-structured learning is used. In contrast to earlier tree-based learning algorithms, LightGBM grows the tree vertically. It is also possible to say that it develops new algorithms at the leaf and tree levels. Benefits of this technique include low memory requirements, GPU learning, and quick processing speed. (Ge, Ke, 2017)

Categorical Boosting: Every new tree constructed by CatBoost and other common gradient boosting implementations is meant to approximate the gradients of the existing model (Dorogush, A.V 2018). Nevertheless, all classical boosting methods suffer from overfitting due to biased pointwise gradient estimates. The gradients utilized at each stage are estimated using the same set of data points that were used to construct the current model. Because it modifies the distribution of expected gradients in any feature space domain in relation to the real distribution of gradients in that domain, this leads to overfitting. When compared to other gradient boosting algorithms, Catboost handles categorical characteristics with ease and outperforms existing publicly available gradient boosting implementations in terms of quality on a variety of well-known publicly accessible datasets (A. Gulin, G. Gusev, 2021). They perform significantly better than current gradient boosting libraries when applied to ensembles of similar sizes.

3.3 DATA PREPARATION

The stratified K-Fold validation approach is utilized to generate the training and testing datasets, as it has demonstrated effectiveness in the field of imbalance data classification on medical data (Khushi, M, 2021).

In order to evaluate the f1 score of our model using an error metric, we divided the dataset into training and testing sets in this study. The former set was used to train the model, while the later set was used to test it. Furthermore, the f1 score received for one individual may change dramatically from that obtained for another test set. Moreover, a set's accuracy might be erroneous since the accuracy obtained for one test set might not match that of another (Khushi, M, 2021). As a result, we employed the stratified K-fold cross-validation technique, which involves splitting the data into folds and using each fold as a testing set periodically.

The performance of classifiers and machine learning models is assessed using train/test splits following the application of the cross-validation approach to minimize biases. Since a greater K suggests less bias toward overestimating the genuine predicted error but also a higher variation and longer running time, a 10-k fold was used for this work. A stratified K-fold cross-validation object is a variation of K-Fold (K=10 in this case) that produces stratified folds. The folds are produced by holding the number of samples for each class constant. It provides train/test indices to further split data into train and test sets.

3.4 EVALUATION METRICS

F1 SCORE

The F1 Score is the weighted average of Precision and Recall. It is used to achieve a balance between the objectives of high accuracy and good recall (Rohit Kundu, 2022). The harmonic mean of recall and precision is used to calculate the F1 Score, which indicates how accurate an exam is. The F1 score range is [0, 1]. It shows our classification's accuracy and robustness (the number of examples in which it is correctly categorized). Consequently, the weighted average of precision and recall may be shown as

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall\ + Precision}$$

F1 Score with "Weighted" Average: It first determines the F1 score for every class separately, averaging them together using a weighted formula depending on the support (number of instances) of each class. This indicates that classes with a higher number of instances affect the average more than classes with a lower number of instances. By assigning greater weight to the classes with more data, it tackles the problem of class imbalance.

3.5 HYPERPARAMETER SEARCH

To introduce new design criteria for the next generation hyperparameter optimization library, Optuna is being employed in this research (Akiba, T., 2019). Optuna outperforms other hyperparameter search techniques for the reasons listed below:

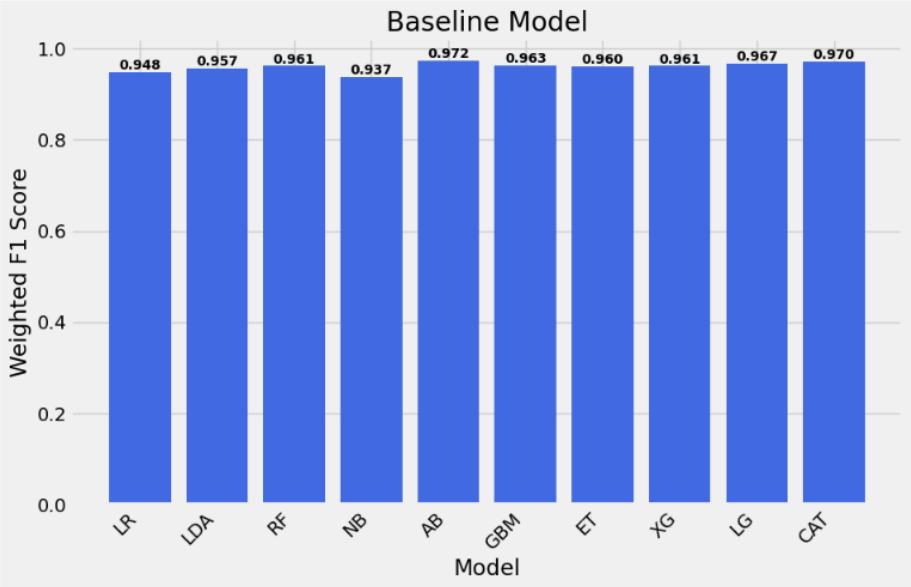
Two key features are the

i. define-by-run API, which enables users to dynamically build the parameter search space
ii. effective application of both searching and pruning algorithms.
iii. A flexible, easy-to-assemble architecture that may be used for a number of tasks, such as scalable distributed computing and lightweight experiments carried out through interactive interfaces

4. RESULT

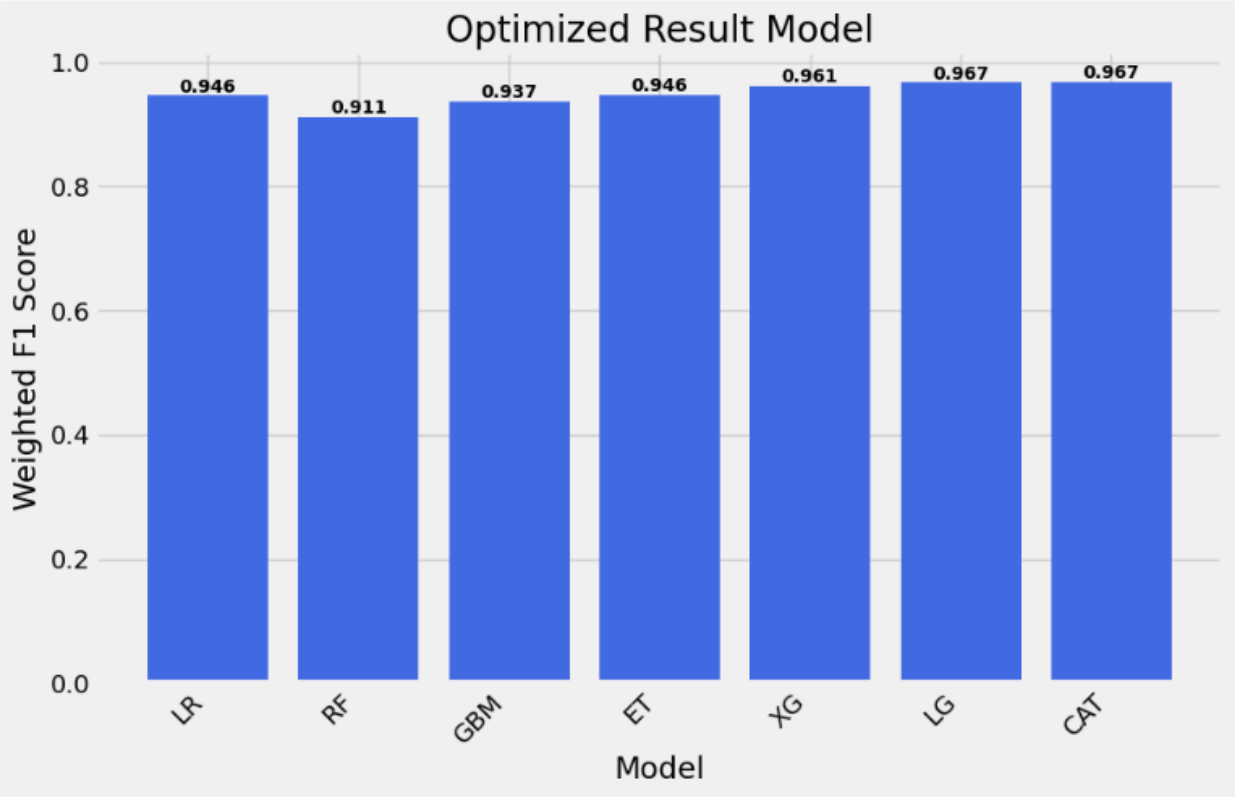
4.1 BASELINE MODEL RESULT

S/N	MODEL	ACCURACY	WEIGHTED F1
1	LOGISTIC REGRESSSION	0.9472	0.9476
2	RANDOM FOREST	0.9612	0.9614
3	GRADIENT BOOSTING MACHINE	0.9629	0.9633
4	EXTRA TREES	0.99594	0.9597
5	XGBOOST	0.9612	0.9614
6	LIGHT GBM	0.9664	0.9671
7	CATEGORICAL BOOSTING	0.9700	0.9703



4.2 OPTIMISED MODEL RESULT

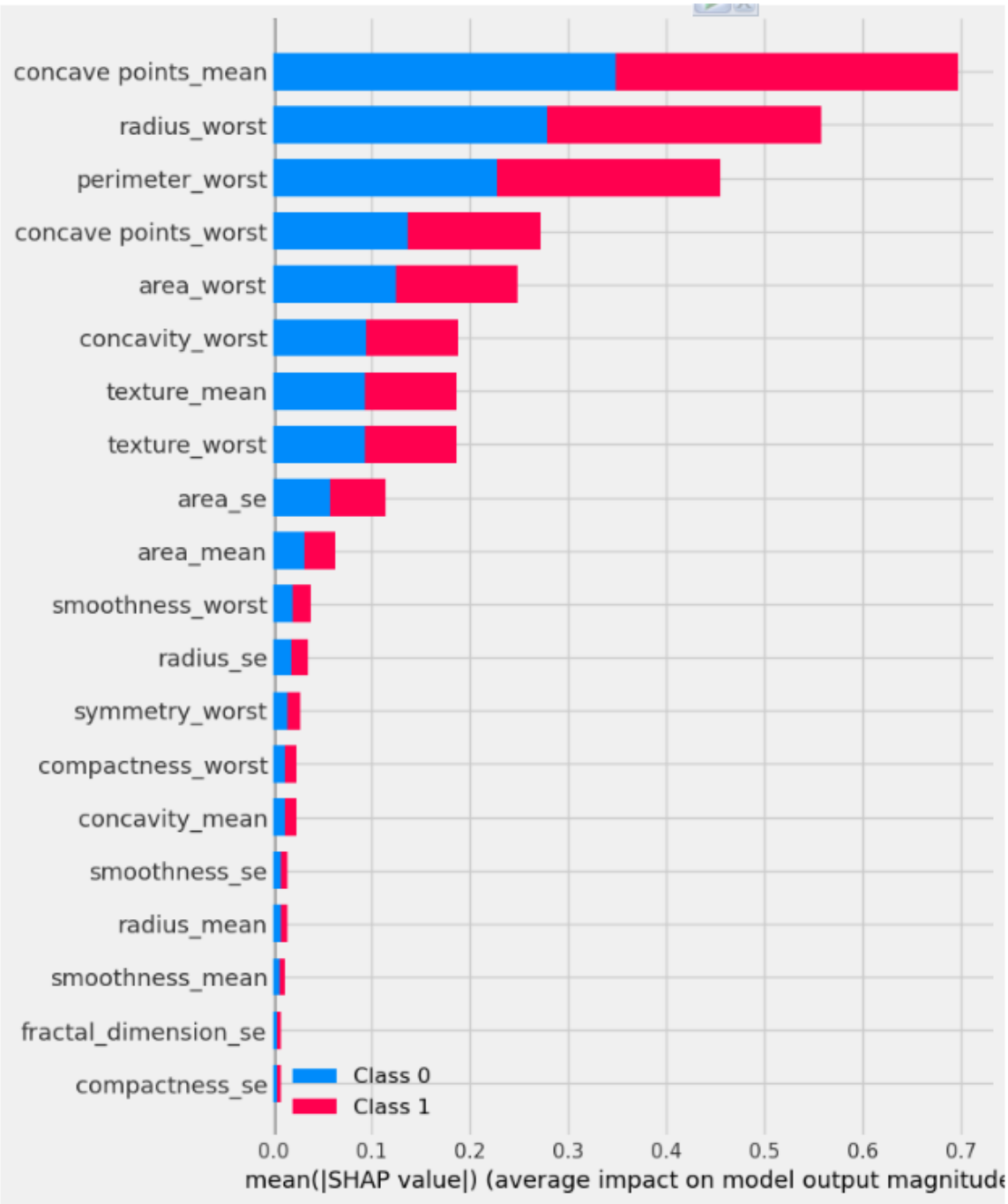
S/N	MODEL	MODEL F1 SCORE	WEIGHTED F1
1	LOGISTIC REGRESSSION	0.9454	0.9459
2	RANDOM FOREST	0.9119	0.9114
3	GRADIENT BOOSTING MACHINE	0.9367	0.9367
4	EXTRA TREES	0.9454	0.9462
4	XGBOOST	0.9612	0.9614
5	LIGHT GBM	0.9665	0.9667
6	CATEGORICAL BOOSTING	0.9665	0.967



4.3 Model interpretability shap analysis

Additionally, we employed the SHapley Additive exPlanation (SHAP) method (Vavilala, et al., 2018) to look into the correlations between the characteristics. To put it briefly, the algorithm is inspired by game theory, which views the interaction between characteristics as a "team" of features, each of which constitutes a member that contributes to the total risk calculation. An instance of the feature interaction records a set of expected values produced by the prediction model. These numbers are inputted into the SHAP algorithm, which then generates a distinct set of values known as "impact values." To determine the probability of hazard and the role of each element separately, the SHAP values provide a dynamic view of the features' interactions.

The SHAP technique also allows one to compare a predicted risk probability for each individual with a baseline prediction, which is the average anticipated probability known as the "base value (Parsa et al, 2020).



Shap Feature Importance for LightGBM Model

5 CONCLUSION

According to the investigation and findings, a machine learning algorithm can reliably classify breast cancer and determine if a patient has the disease or not. The investigation found that not all algorithms were able to improve after optimization; this can be attributed to the small sample space of data used to train the algorithm, but some algorithms did improve. To make better use of this outcome, deploying the best performing algorithm to the cloud would require the assistance of a surgical oncologist and allow for cross-hospital use of the results.

6 REFERENCES

1. Armengol, E. and De Mantaras, R.L., 1998. Machine learning from examples: Inductive and Lazy methods. Data & Knowledge Engineering, 25(1-2), pp.99-123.

2. A. Gulin, G. Gusev, L. Ostroumova Prokhorenkova, A. V. Dorogush, and A. Vorobev. Catboost: unbiased boosting with categorical features. arXiv preprint arXiv:1706.09516, 2017.

3. Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M., 2019, July. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).

4. Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, pp.1937-1967.

5. Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

6. Bronshtein Adi, “Pandas” Python Library’., (2019), Towards Data Science: [Online] [Accessed https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library- f1b678f34673.

7. D. Delen, "Analysis of cancer data: A data mining approach", Expert Syst., vol. 26, no. 1, pp. 100-112, Feb. 2009.

8. Dorogush, A.V., Ershov, V. and Gulin, A., 2018. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363.

9. E. Frank and I. H. Witten, "Data mining: Practical machine learning tools and techniques with Java implementations", ACM SIGMOD Rec., vol. 31, no. 1, pp. 76-77, Mar. 2005.

10. GCO. Number of New Cases in 2019, Both Sexes, All Ages. Available online at: http://gco.iarc.fr/today/data/factsheets/populations/900-world-factsheets.pdf.

11. GeeksforGeeks, 2021, NumPy in Python.. [Online] [Accessed https://www.geeksforgeeks.org/numpy-in-python-set-1-introduction

12. Hackeling, G., 2017. Mastering Machine Learning with scikit-learn. Packt Publishing Ltd

13. Huang, B. Cai, Z., Gu, Q., and Zhu, L.,(2008). Data Mining on Imbalanced Data Sets. In 2008 International Conference on Advanced Computer Theory and Engineering (pp. 1020-1024). doi:10.1109/ICACTE.2008.26

14. ISTAT Data. Available online at: www.demo.istat.it ISTC. Data Available online at: www.demo.istat.it

15. J. Brownlee (2021), "Scikit-Learn: A Python Machine Learning Library.", Machine Learning Mastery.[Online],[Accessed https://machinelearningmastery.com/a-gentle-introduction-to- scikit-learn-a-python-machine-learning-library.

16. Machado, M.R., Karray, S. and de Sousa, I.T., 2019, August. LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. In 2019 14th International Conference on Computer Science & Education (ICCSE) (pp. 1111-1116). IEEE

17. M. Bahaj and Y. Khourdifi , "Applying best machine learning algorithms for breast cancer prediction and classification", Proc. Int. Conf. Electron. Control Optim. Comput. Sci. (ICECOCS), pp. 1-5, Dec. 2018.

18. M. Shaheen, , S. Faruq, M. Shahbaz and S. A. Masood, "Cancer diagnosis using data mining technology", Life Sci. J., vol. 9, no. 1, pp. 308-313, 2012.

19. P. Chandra and M. K. Gupta, "A comprehensive survey of data mining", Int. J. Inf. Technol., pp. 1-15, Feb. 2020.

20. Rohit Kundu (2022), F1 Score in Machine Learning: Intro & Calculation. V7: [Online] [Accessed] https://www.v7labs.com/blog/f1-score-guide.

21. Salod, Z. and Y. Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol", J. Public Health Res., vol. 8, no. 3, pp. 1677, Dec. 2019.

22. Tin Kam Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 278-282 vol.1, doi: 10.1109/ICDAR.1995.598994.

23. Vachanaram, A. R Mora, E., Varughese, , Nardin, S., F. M., D'Avanzo, F.,., Rossi, V.,Gennari, A. (2020). Breast Cancer Survivorship, Quality of Life, and Late Toxicities. Frontiers in Oncology, 10. doi:10.3389/fonc.2020.00864

24. Vavilala, M. S., Lundberg, S. M., Nair, B., Horibe, M., Eisses, M. J., Adams, T., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat. Biomed. Eng. 2, 749–760. doi: 10.1038/s41551-018-0304-0

25. Virtanen, P., Gommers, R., Oliphant, T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2